LENDING CLUB CASE STUDY

SUBMISSION

Submitted By:

1. **Rajeev Agarwal**
2. **Pradeep Bhaganna**
3. **Vivek Mallampalli**

Lending club case study

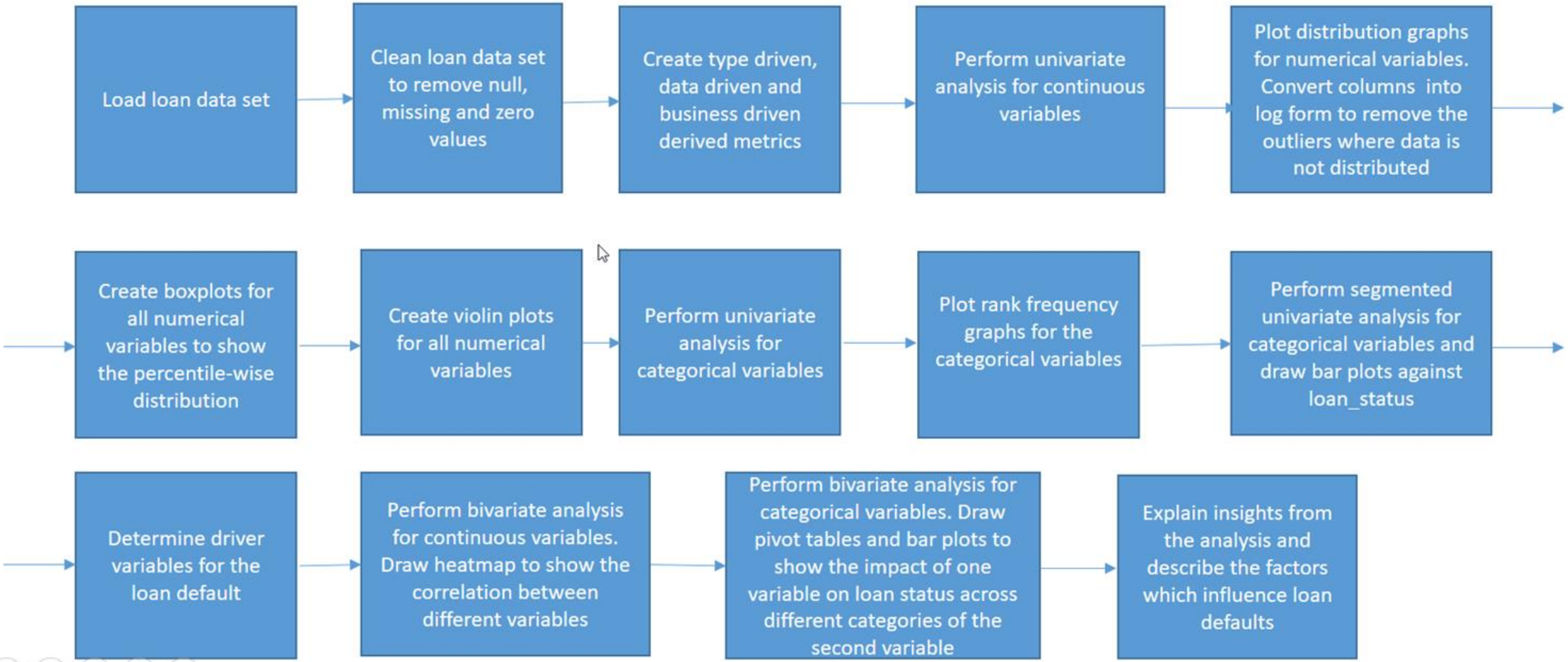
Business objective: The objective is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. Lending club company wants to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

Goals of data analysis:

- Univariate and Segmented Univariate Analysis - Identify at least 5 important driver variables (i.e. variables which are strong indicators of default).
- Bivariate Analysis - Identify the important combinations of driver variables. The combinations of variables are chosen such that they make business or analytical sense.
- Explain the most useful insights from the given data.

Problem solving methodology

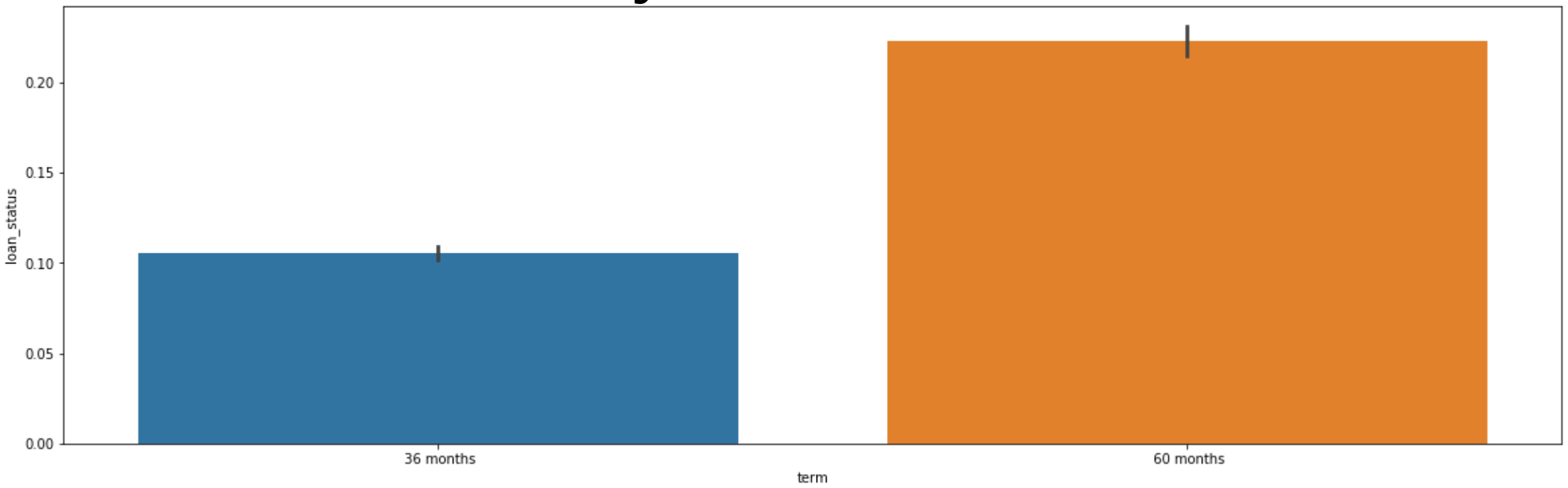
Flow Chart



Data Cleaning and Understanding

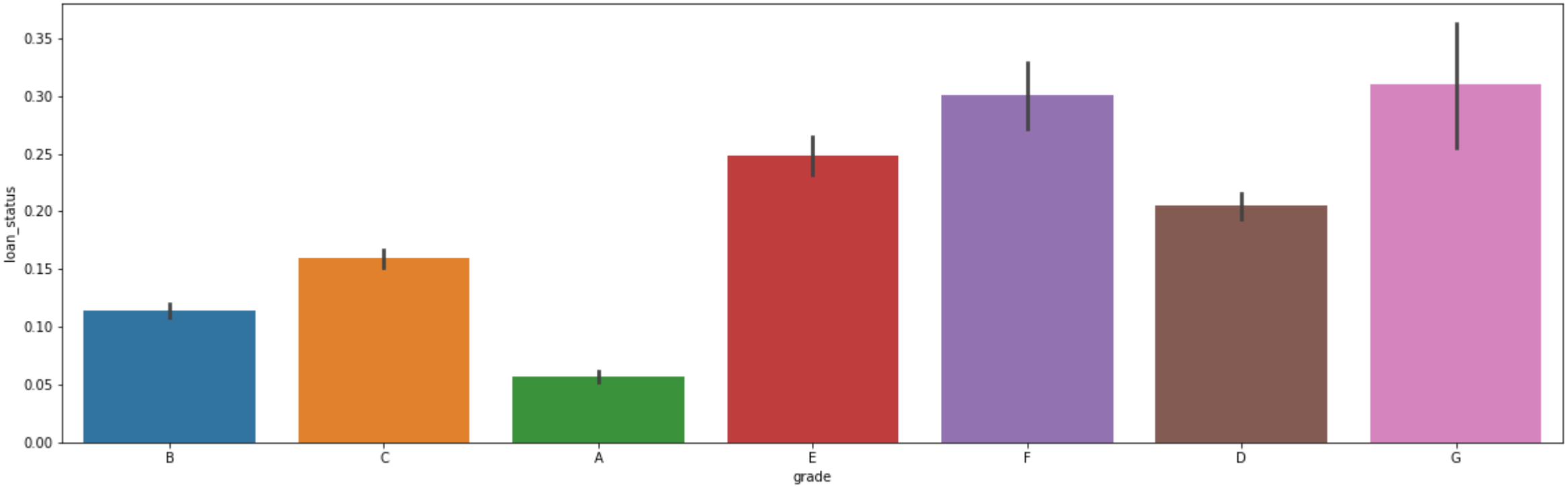
- We have taken complete loan data for all loans issued through the time period 2007 to 2011.
- There is no transactional history of the applicants whom loan was rejected and so this data is not available with the company (and thus in this dataset).
- Following data set was used to do the analysis:
  - Loan Data set - loan.csv
- Data_Dictionary spreadsheet was used to understand the metadata.
- Default encoding was used to load the data set.
- An overwhelming amount of data and at times incomplete data drives the need of data cleaning to pinpoint which variables are of particular importance towards the business objectives
- Data was cleaned before analysis to remove null and invalid values.
  - Columns having most of the null values were dropped.
  - For columns where we had very few number of null values, we deleted the rows having the null values.
  - Columns where the values were constant in all rows were dropped.

Univariate Analysis - Term



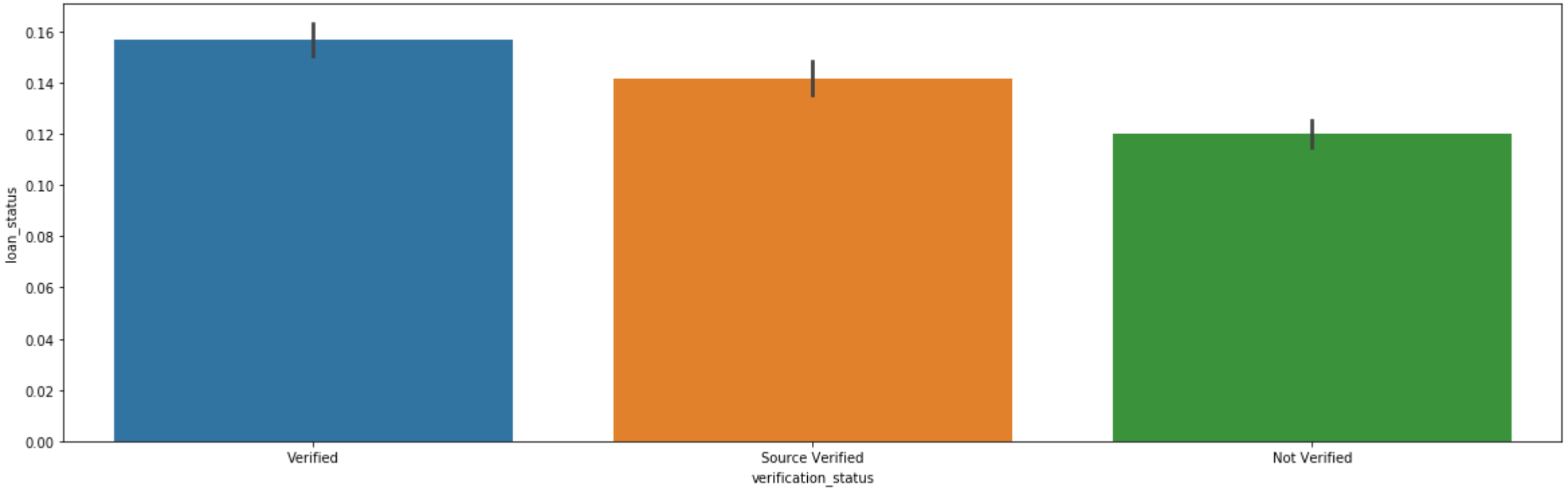
- Loans with 60 month terms are over 10% more likely to default on payment than those with 36 month terms
- Additional scrutiny can be added to evaluation processes surrounding 60 month term loans

Univariate Analysis - Grade



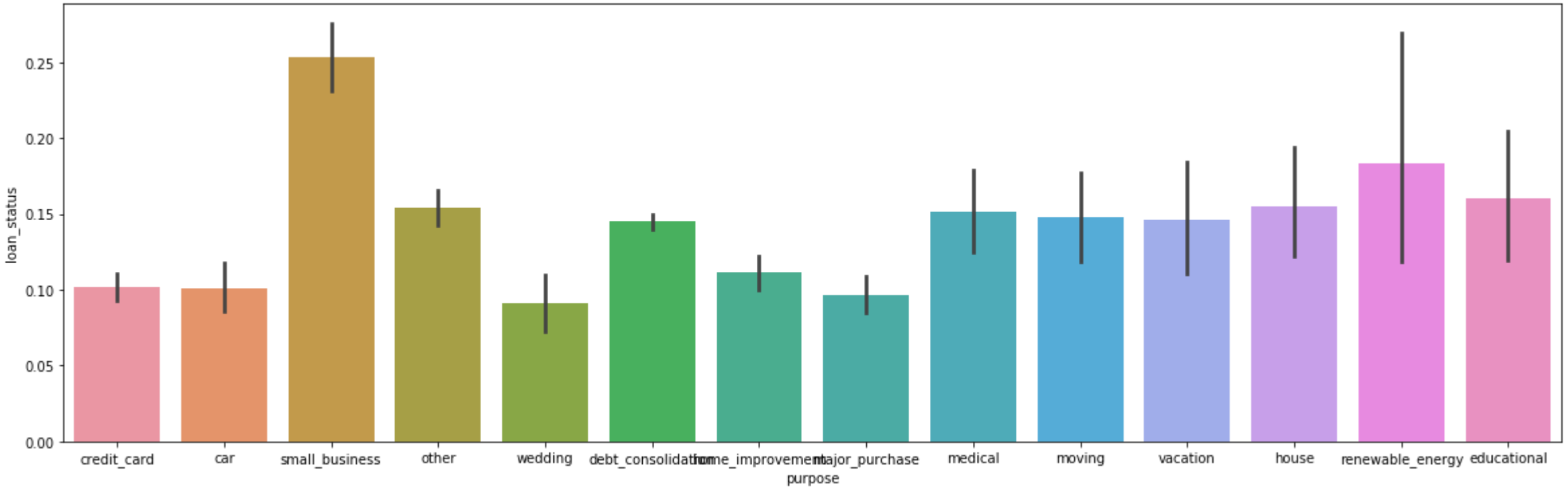
- Grade A loans have the lowest default rate and default rates increase gradually across the grade levels
- This is as expected because grade A loans are rated as least risky and grade G loans are most risky, these applicants are probably assigned higher interest rates

Univariate Analysis - Verification Status



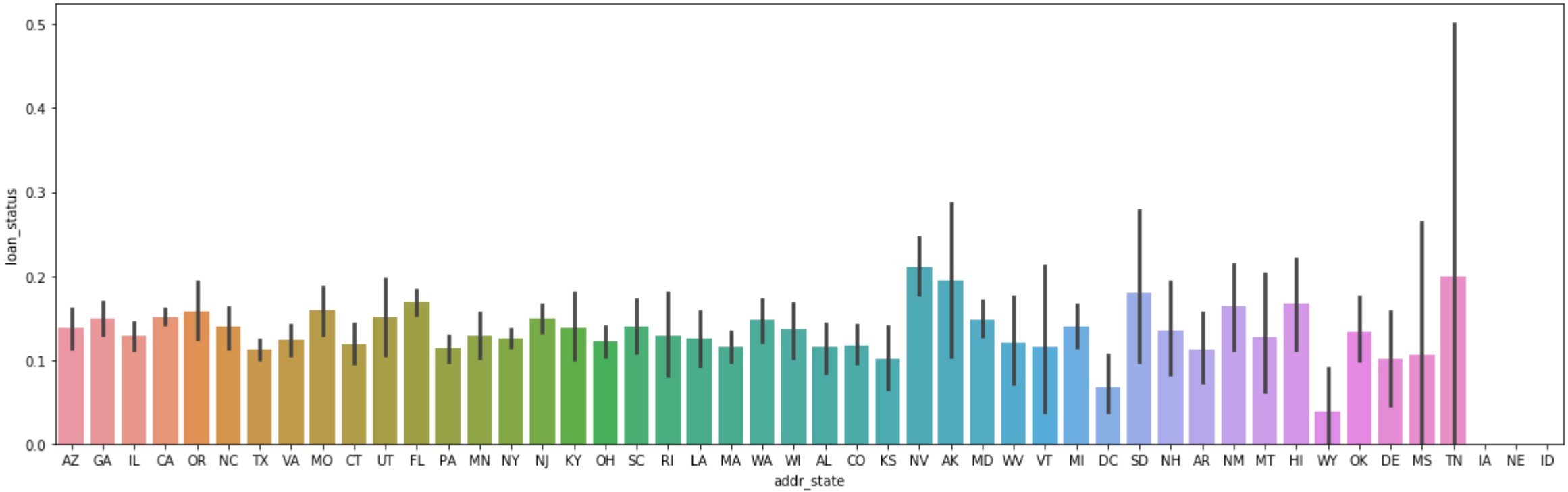
- Verified loans have the highest default rates, followed by loans having their sources verified, and finally those loans which have not been verified
- This is of particular interest because we would expect verified loans to have the lowest default rates
- This indicates that the business must rework its verification processes to fix its inefficiencies in this realm

Univariate Analysis - Purpose



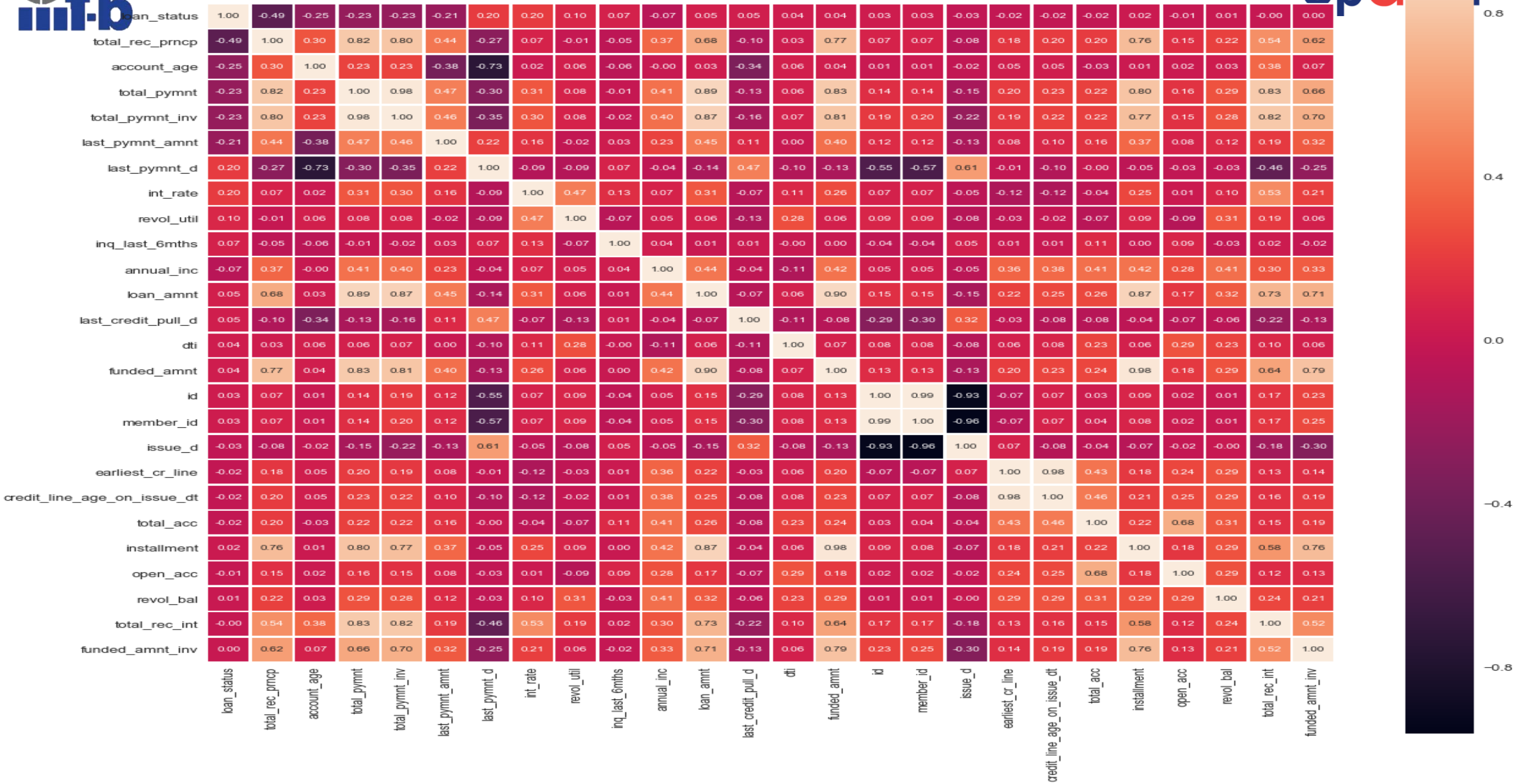
- The purpose that has the lowest default rate is wedding which is not unusual because this decision usually takes place if both individuals are financially secure
- Small business loans have the largest default rate which is not surprising considering the large amount of risk involved with start-up companies
- Renewable energy is next in highest default rates which is probably due to the inherent risk involved in an industry that is up and coming
- Educational loans are the next highest which is reasonable considering education is the largest debt facing Americans today, surpassing even mortgage debt

Univariate Analysis - Address State



- Nevada (NV) has the highest level of default rate followed by Alaska (AK) and Tennessee (TN) these states might require more scrutiny when evaluating
- Wyoming (WY), District of Columbia (DC), and Kansas (KS) have the lowest level of default rates as compared with the others, they can be considered less relatively risky

Correlation Heat Map of Numerical Variables



Correlation Heat Map Explained

Continuous Variables with strong negative correlation with loan default (where loan_status:0=Not default,1=Default):

- total_rec_prncp (-0.49) : Loans with higher principal received to-date are less likely to default.
- account_age(-0.25) : Loans with higher account age are less likely to default. Newer loans have more probability to get defaulted.
- total_pymnt and total_pymnt_inv (-0.23) : Loans with higher total payment received are less likely to default.
- last_pymnt_amnt (-0.21) : Loans where last payment amount was higher are less likely to default.

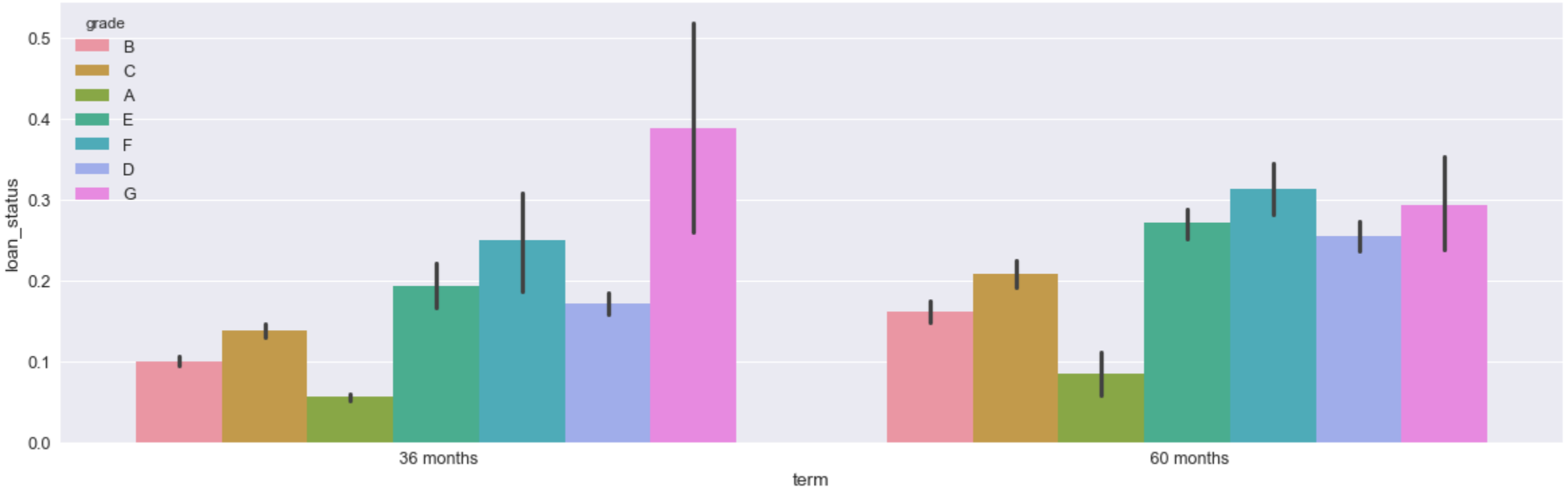
Continuous Variables with strong positive correlation with loan default (where loan_status:0=Not default,1=Default):

- int_rate (0.20) : Loans with higher interest rate are more likely to default.
- revol_util (0.10) : Loans from borrower's with higher revolving line utilization rate are more likely to default.

Other important bivariate correlations:

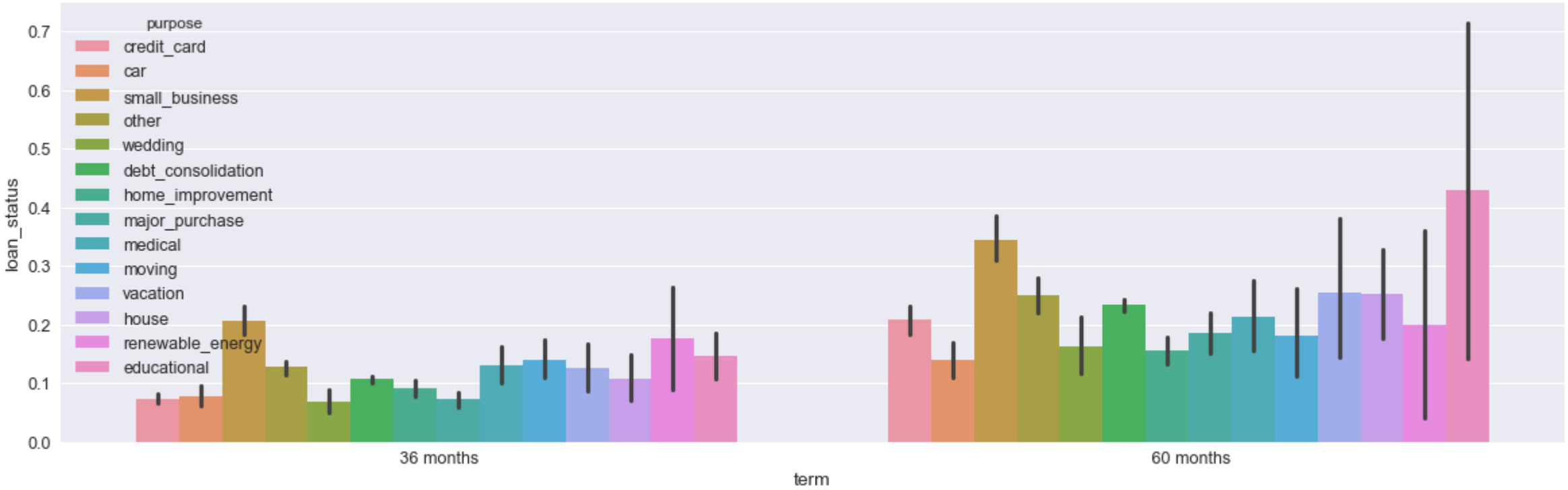
- total_pymnt and total_pymnt_inv have correlation coeff value 0.98, so one variable can be ignored from the analysis.
- total_recv_prncp has strong correlation with total_pymnt, total_pymnt_inv, funded_amnt and installment variables.
- total_pymnt and total_pymnt_inv have strong correlation with loan_amnt, funded_amnt, installment and total_rec_int variables.
- int_rate has strong correlation with revol_util and total_rec_int variables.
- dti has strong correlation with open_acc, revol_util, revol_bal, total_acc, int_rate variables.
- funded_amnt has strong correlation with loan_amnt, installment, total_pymnt, total_pymnt_inv, funded_amnt_inv, total_recv_prncp, annual_inc, total_recv_int.

Bivariate Analysis - Term and Grade



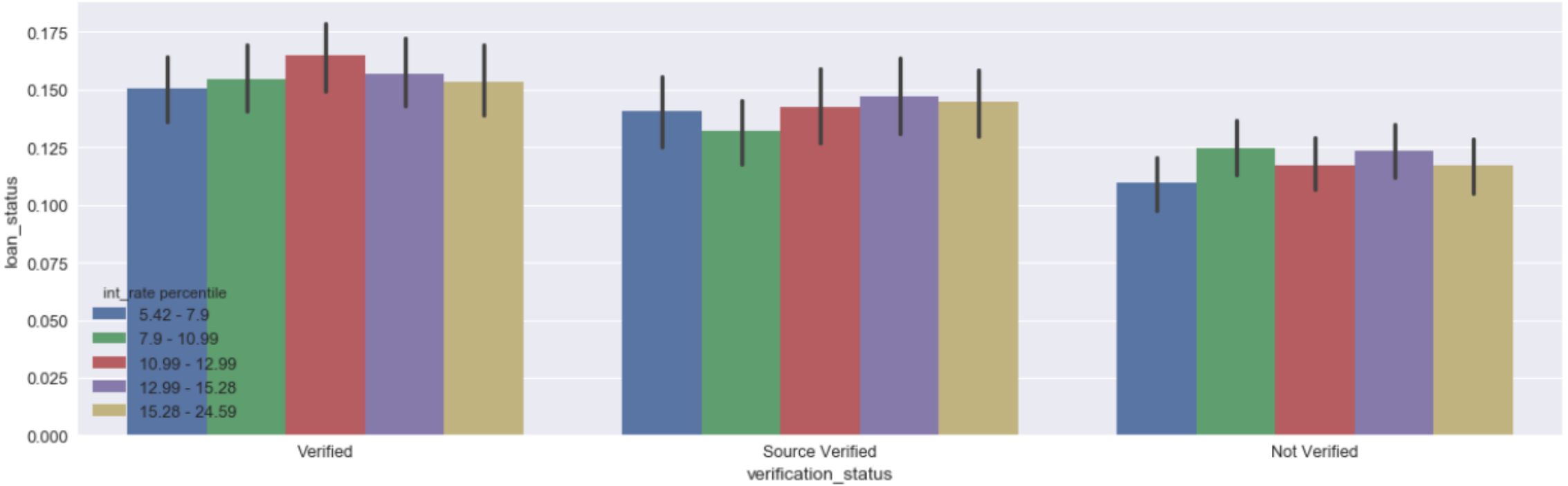
- Within 36 month loans, grade G loans are driving the default rates exceeding the next highest grade E by over 10%
- Within 60 month loans, Grades D-G are closer to each other in default rates and together drive default rates for the term group
- Expected trends show throughout in that default rate increase across grades in both term groups for the exception of the slight drop in defaults for grade G in 60 months making grade F the highest there

Bivariate Analysis - Term and Purpose



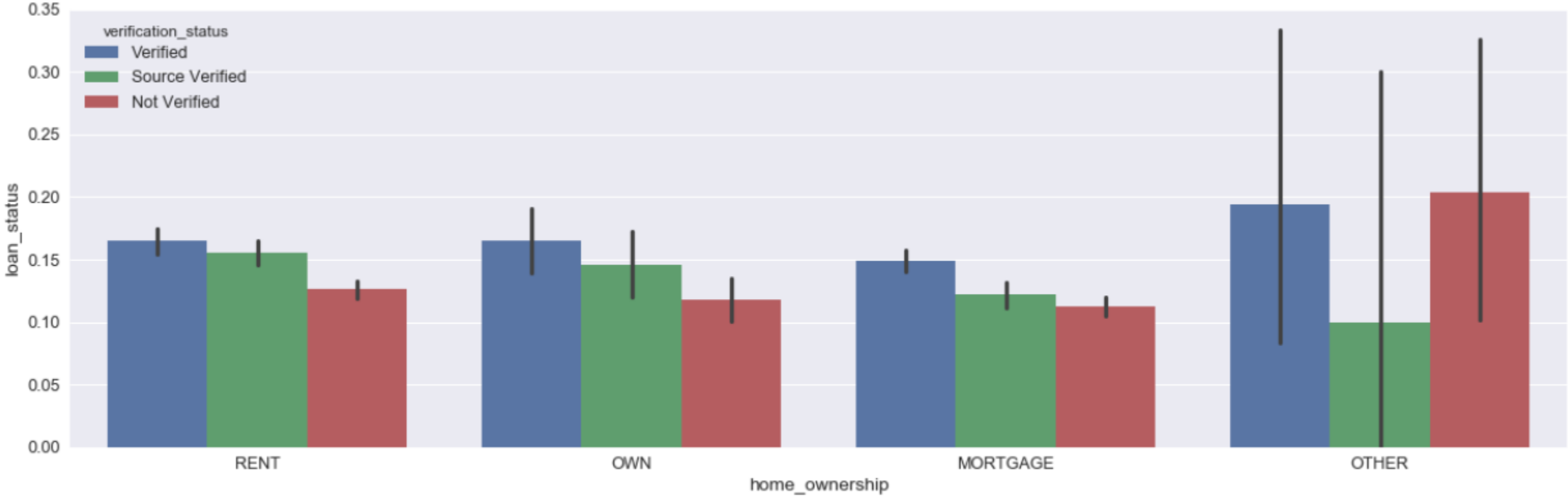
- In 60 month loans, educational loans carry the largest default rates which is as expected because educational loans are usually longer term and as stated earlier are the largest source of debt for Americans
- In 30 month loans, small business loans lead since most educational loans are not shorter term and because of the inherent risk of startups. They are the second largest default group in the 60 month loans

Bivariate Analysis - Verification Status and Interest Rate Percentile



- Corresponding interest rate percentiles get lower across the different verification statuses
- Under verified, we can see that the interval containing the median has the highest default rate indicating that an average individual in this status of loans would most likely default compared to the other statuses
- This indicates again that the verification processes need to be reworked because we would expect not verified incomes to have the larger default rates because of the lack of scrutiny taken before approval

Bivariate Analysis - Home Ownership and Verification Status



- Those in the Rent and Own statuses have similar default rates, but those in the Mortgage Category have lower rates
- Because this data is post-2007 after the burst of the housing bubble, banks were more stringent on whom they gave loans to, thus people who had mortgages were less likely to default on any loan

Conclusions

- 60 month term loans are most likely to default by over 10%
- Lower grade loan can be applied with higher interest rates than they are at now to discourage potential borrowers who are likely to default
- The income verification process will need to be reworked because those who were not verified are actually less likely to default which is not what would be expected
- Some states such as NV, AK, and TN have higher risk of default, these states deserve more scrutiny in approving loans from applicants who live in them
- To prevent the large default rate caused by educational loans in the 60 month term category, cosigners or an additional borrower can be made required since most students are not financially stable enough to pay off their own debt
- Small businesses should be required to go through a more stringent approval process to ensure lower default rates such as collateral
- Rent and Own home ownership statuses should have more stringent approval statuses similar to Mortgage status